STAT 840: Final Project

Mark Chintis

May 11, 2019

I. Introduction

The College Board's Standarized Achievement Test (SAT) is a test that every student planning to go to college in the United States must take. For students, it can be the most rigorous and stressful period in their high school careers. For school administration and teachers, the SAT can be a meaningful indicator of overall student academic performance. Test results can be a useful resource, asset and influence school practices provided there are methods of interpretting them. The administration at Santa Cruz Cooperative School (SCCS) expressed an interest in improving the interpration of their students' SAT scores. This study hopes to provide one such method by using past student grades and other academic measures within the school's control to create a linear regression model to be used for the prediction of future SAT scores at the school. SCCS administration requested that the study be focused on the SAT mathematics section scores and thus only mathematics related predicator variables are included in the study design.

A. Study Design.

The collection of student math course scores from eighth grade through eleventh grade and Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) test scores from eighth grade through tenth grade. These metrics are used to assess variation and predict future SAT scores. Students at SCCS are placed into either a regular track or an advanced track. Regular track individuals are coded as 0 and advanced track individuals are coded as 1. All metrics were recorded from 2010 to 2018, however only 97 complete cases were considered for analysis. Table 1 displays the first 6 observations of this data set. Scores from the eighth, ninth and tenth grade MAP test are the variables **MAP8**, **MAP9**, **MAP10**, respectively. These MAP scores are based on the RIT score developed by NWEA, and can range from 180 to 285. **SAT** scores are standard from the College Board and are scaled from 200 to 800. The other variables are: **Math8**, **Math9**, **Math10**, and **Math11** each correspond to a student's end of course grade on a 0 to 100 scale.

Table 1: Project Data Set

Track	MAP8	Math8	MAP9	Math9	MAP10	Math10	Math11	SAT
1	243	84	246	89	255	85	90	640
1	236	88	240	88	245	82	92	700
1	236	79	238	76	250	70	84	700
1	244	84	249	89	257	87	91	660
0	230	93	239	82	244	78	78	540
1	251	85	252	96	268	98	98	750

B. Aims.

The primary objective of this study is to choose the regression model of selected variables that best describes the variation of SAT scores. The model selected will be used by the school administration to predict future students' SAT scores.

C. Statistical Model.

A multiple regression model is considered in this study. Let

Y_i = SAT math score
X_{i1}= track of mathematics courses (either regular or advanced) for the i^{th} student
X_{i2}= 8^{th} grade MAP score for the i^{th} student
X_{i3}= 8^{th} grade math course score for the i^{th} student
X_{i4}= 9^{th} grade MAP score for the i^{th} student
X_{i5}= 9^{th} grade math course score for the i^{th} student
X_{i6}= 10^{th} grade MAP score for the i^{th} student
X_{i7}= 10^{th} grade math course score for the i^{th} student
X_{i8}= 11^{th} grade math course score for the i^{th} student

The initial model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + \varepsilon_i$$

where $\varepsilon_i \sim iidN(0, \sigma^2)$, $i = 1, 2, ..., 97$, and $\beta_0, \beta_1, ..., \beta_8$, and σ^2 are the unknown model parameters.

II. Preliminary Analyses.

A. Bivariate Associations.

Figure 1 gives a scatterplot matrix used to indicate any linear association between all variables. Pearson coefficients are overlaid on the scatterplot matrix for convenience. We see from the scatterplots and the Pearson coefficients that there are strong linear associations between most of the predictor variables. Of particular strength are the associations between 8^{th},9^{th},10^{th}, and 11^{th} grade math scores.

B. Screening of Covariates and Verification of Assumptions

From the results of automatic variable selection methods along with criterion-based statistics, X_2 through X_5 and X_7 were removed from the model. Variable X_1 is looked into as a possible addition to our model and is included as a part of a full model for hypothesis testing. Partial residual plots, residual-versus-fitted plots, and measures of influence were investigated and had only few high influence points and no issues with linearity, constant variance, independence, or normality were identified.

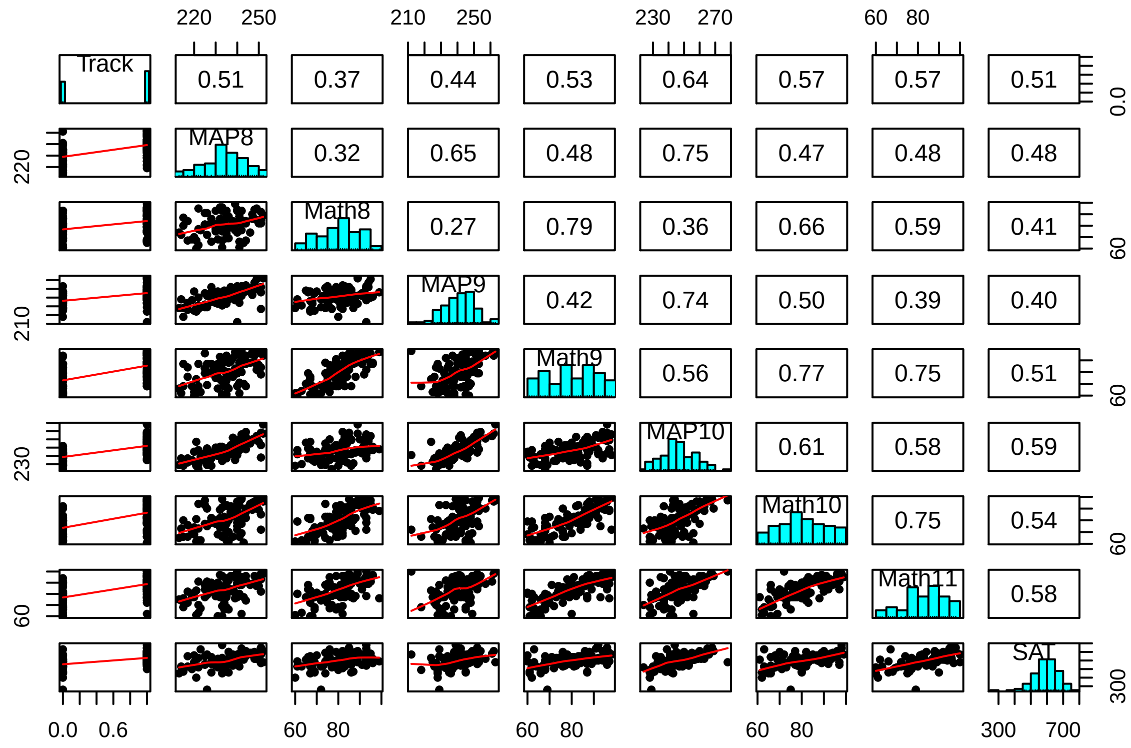


Figure 1: Scatterplot and Correlation Matrix

C. Final Model

The final model is given by

$$Y_i = \beta_0 + \beta_6 X_{i6} + \beta_8 X_{i8} + \varepsilon_i$$

where $\varepsilon_i \sim iidN(0, \sigma^2)$, $i = 1, 2, ..., 95$, and $\beta_0, \beta_6, \beta_8$, and σ^2 are the unknown model parameters.

III. Statistical Analysis.

Through automatic variable selection techniques described in the Appendix, two models are considered for hypothesis testing and validation. The full model for the hypothesis tests is given by

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_6 X_{i6} + \beta_8 X_{i8} + \varepsilon_i.$$

The reduced model is given by

$$Y_i = \beta_0 + \beta_6 X_{i6} + \beta_8 X_{i8} + \varepsilon_i.$$

We wish to test the following hypotheses:

$$
\begin{aligned}
H_0 :&\quad \beta_1 = 0\\
H_1 :&\quad \beta_1 \neq 0.
\end{aligned}
$$

The general linear test approach gives the decision rule to reject the null hypothesis that $\beta_1 = 0$, if $F^* = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F} > F_{.95,1,91}$. Since $F^* = 0.32 < F_{.95,1,91} = 3.946$, we fail to reject the null hypothesis that $\beta_1 = 0$, thus we conclude that the incorporating the variable X_1 into the model is unnecessary. Leading us to a final model of that of the reduced model above.

```
## Analysis of Variance Table
##
## Model 1: Y ~ X6 + X8
## Model 2: Y ~ X1 + X6 + X8
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     92 251013
## 2     91 250135  1    878.67 0.3197 0.5732
```

The fitted model is displayed below. The rate increase of SAT scores is, on average, 2.772 (95% CI 1.517, 4.027) for every one point increase in 10^{th} grade MAP test score. The 10^{th} grade MAP test score accounts for 17.3% while 11^{th} grade course score explains 19.1%.

```
##
## Call:
## lm(formula = Y ~ X6 + X8, data = project_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.001  -38.358    0.386   33.008  200.035
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -340.8389   133.0773  -2.561   0.0121 *
## X6             2.7717     0.6320   4.386 3.07e-05 ***
## X8             3.1098     0.6666   4.665 1.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.23 on 92 degrees of freedom
## Multiple R-squared:  0.504,  Adjusted R-squared:  0.4933
## F-statistic: 46.75 on 2 and 92 DF,  p-value: 9.781e-15
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X6         1 195720  195720  71.734 3.768e-13 ***
## X8         1  59386   59386  21.766 1.042e-05 ***
## Residuals 92 251013    2728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                  2.5 %     97.5 %
## (Intercept) -605.141927 -76.535828
## X6             1.516614   4.026882
## X8             1.785941   4.433655
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X6         1  40352   40352  14.310 0.0006420 ***
## X8         1  42028   42028  14.904 0.0005173 ***
## Residuals 32  90237    2820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To validate the predition capability of our model, a random sample of 35 observations from the project data set are chosen to comprise a validation set. The observed mean square prediction error $MSPE = 2819.895$ (shown above). This is very close to the $MSE = 2728.404$, therefore giving an appropriate, unbiased indication of the predictive ability of the model.

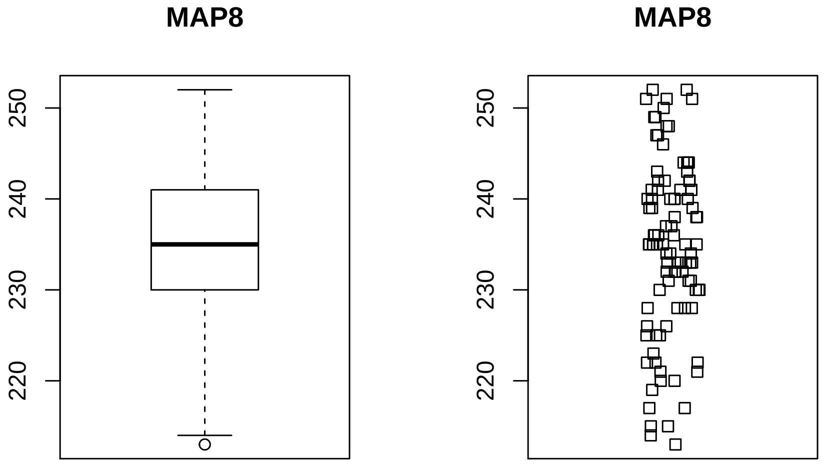
IV. Summary of Findings.

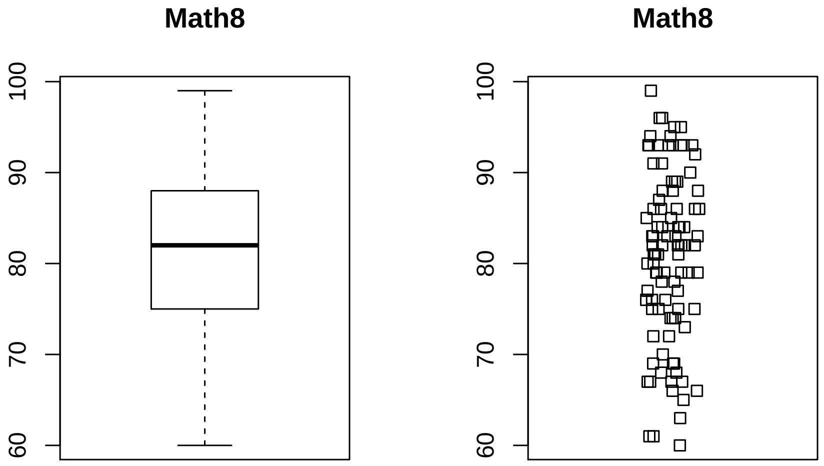
The final model seems to give a fairly strong linear relationship between the variables for a student's 10^{th} MAP score and 11^{th} grade math course score. Although this report shows that predictions can be done, it should be taken with caution. Nearly 50 percent of the variation in SAT scores is still unaccounted for. The limitations of this study are the predictor variables of an SAT score. There are many other variables that can lead to variation in scores that is uncontrollable and difficult to measure by our school administrations. Factors such as outside tutors, amount of SAT specific practice, family emergencies, etc. can play a large role in exlaining the unaccounted variation.

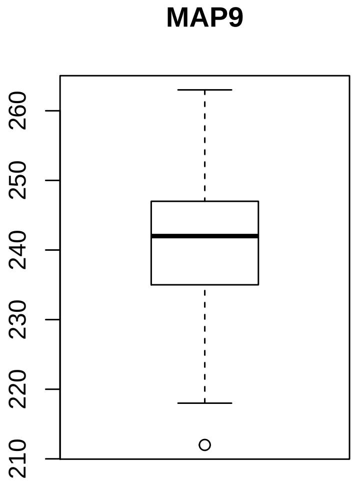
V. Appendix

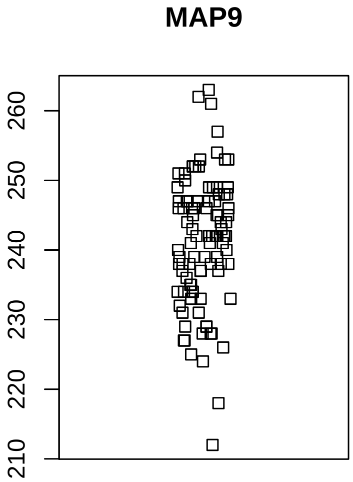
A. Diagnostics of Predictors

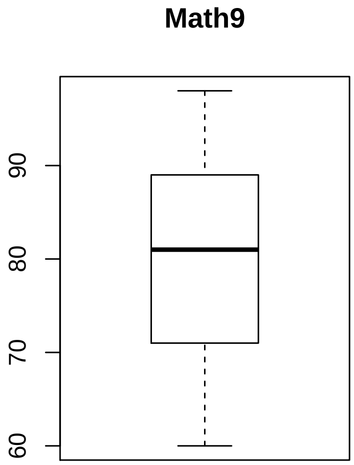
Boxplots and jittered strip plots of each predictor variable are prepared in a side by side comparison in the figures below. The track variable is a categorical variable so we view the distribution with a histogram. All variables seem near normal distributed with no major outlier concerns.

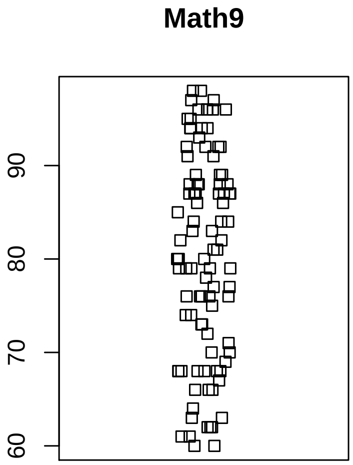

MAP8
220
230
240
250
MAP8
220
230
240
250

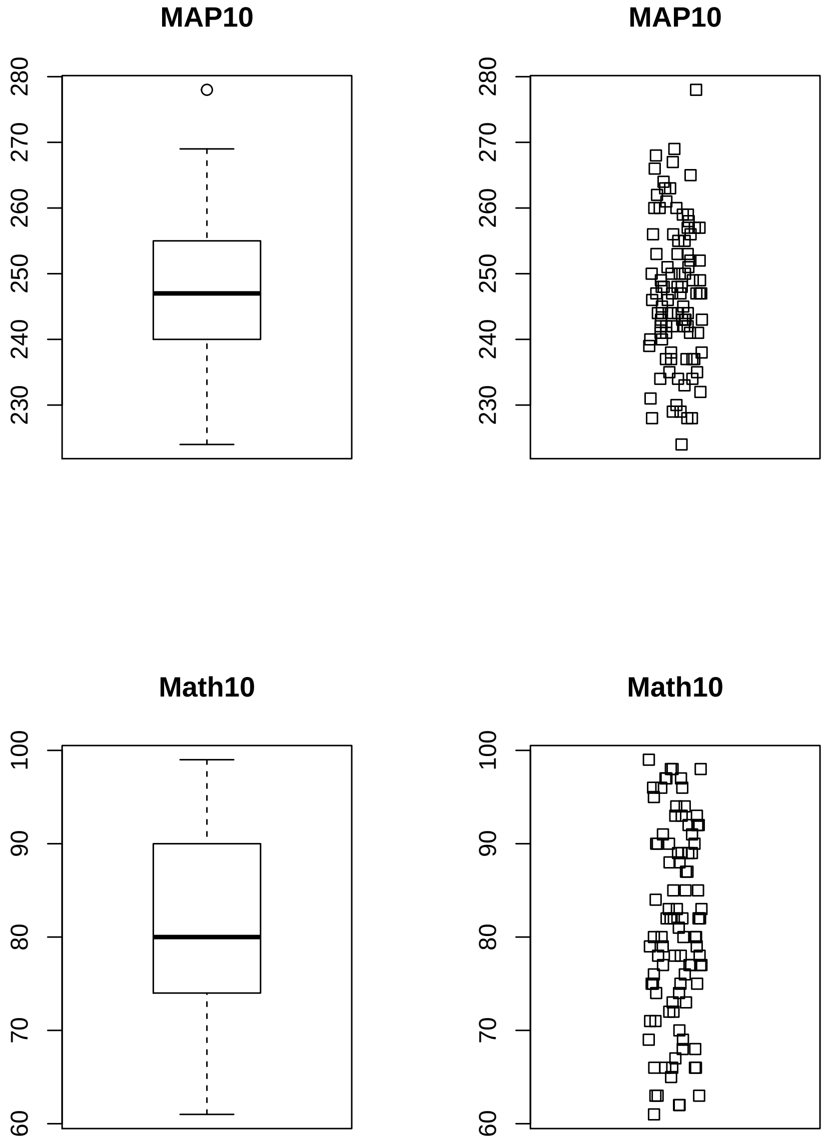

Math8
60
70
80
90
100
Math8
60
70
80
90
100

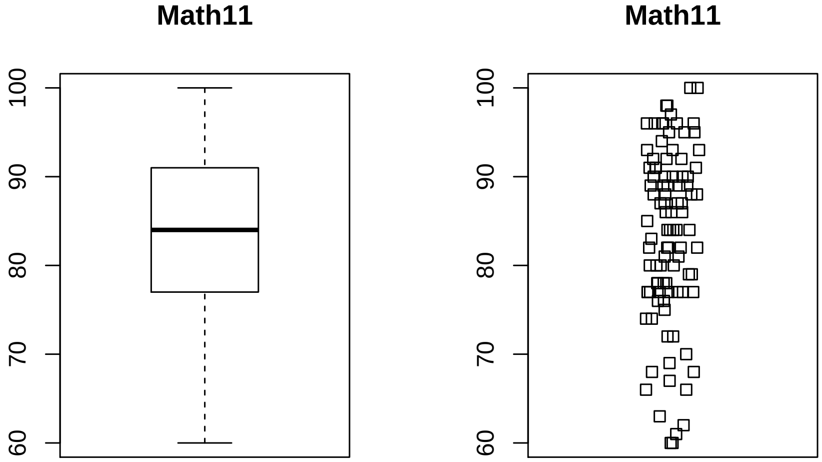

MAP9
260
250
240
230
220
210

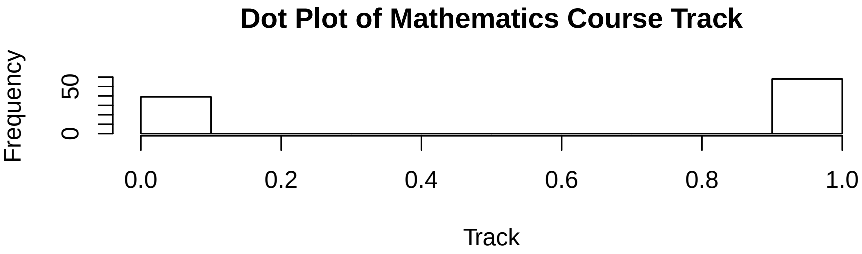

MAP9
260
250
240
230
220
210


Math9
90
80
70
60


Math9
90
80
70
60

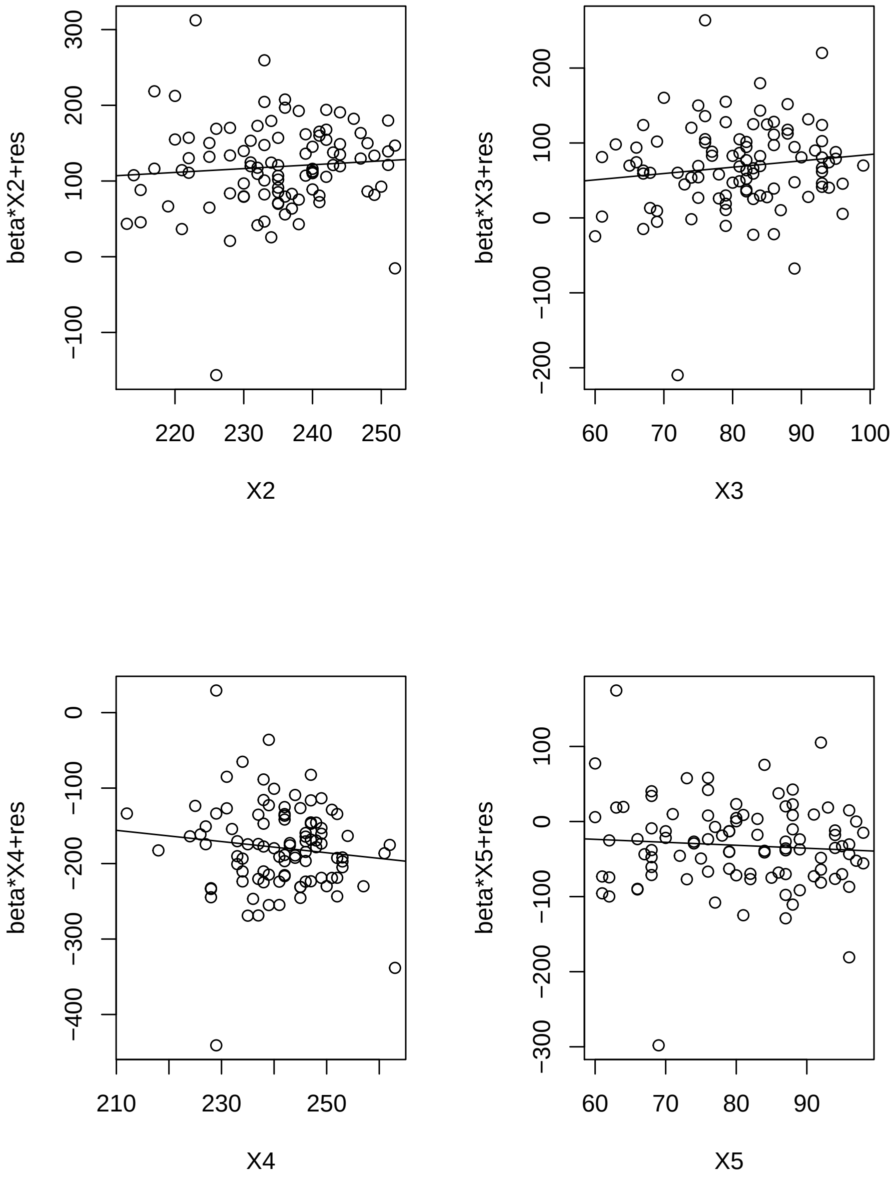

MAP10
280
270
260
250
240
230
MAP10
280
270
260
250
240
230
Math10
100
90
80
70
60
Math10
100
90
80
70
60

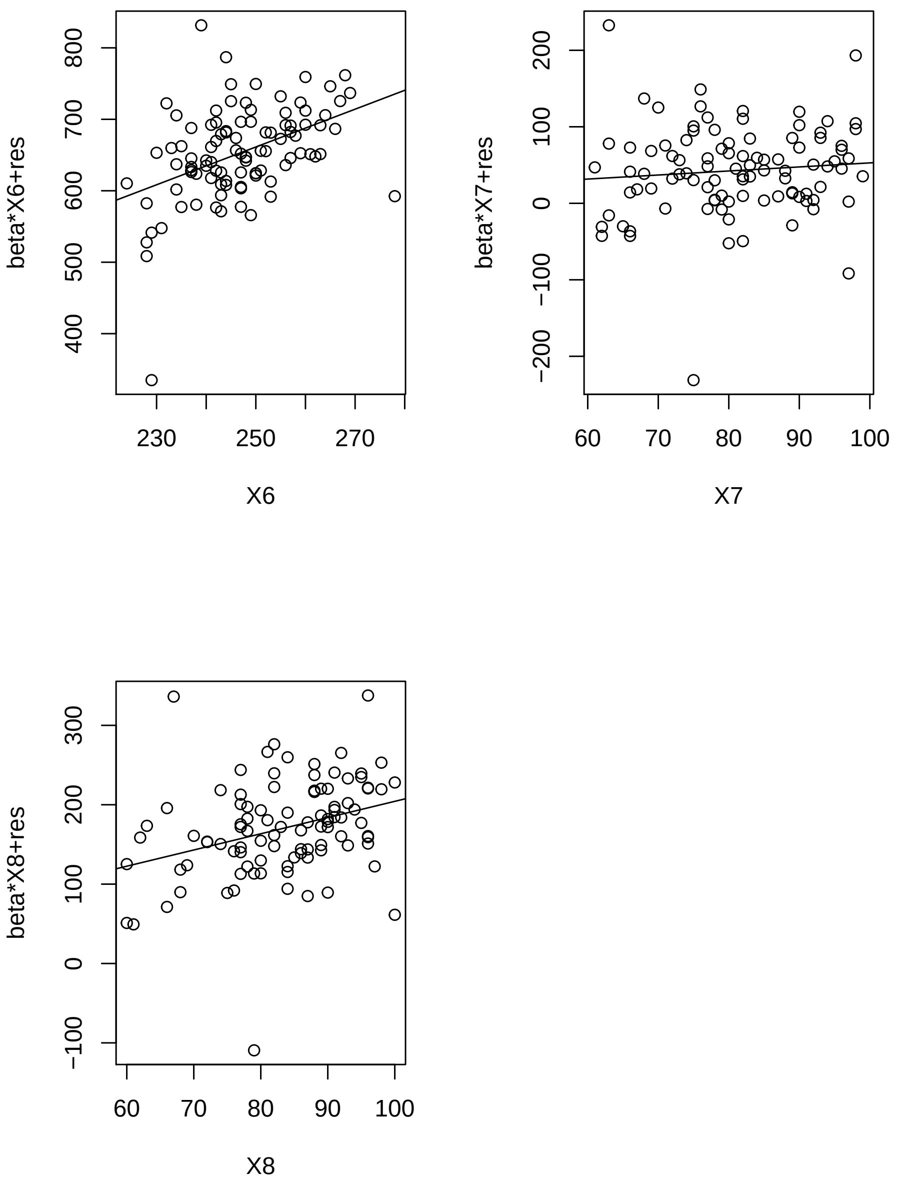




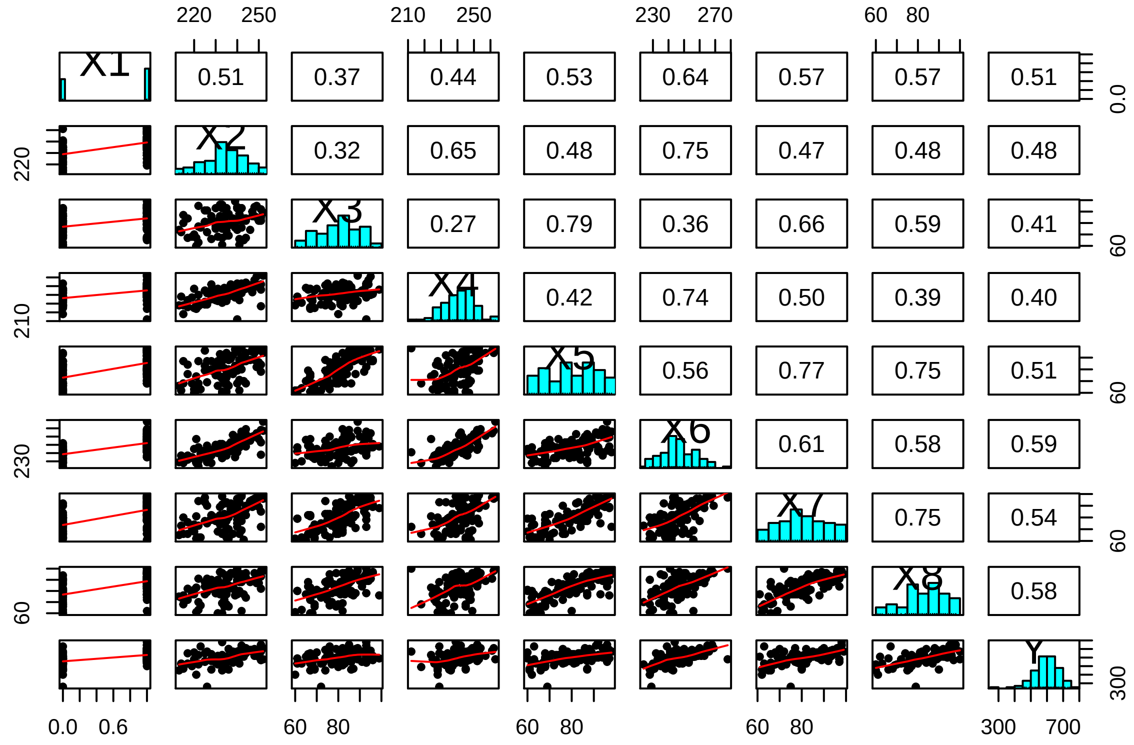
B. Screening of Predictors

1. Below are added variable plots for each covariate. From these plots we can begin to establish that some variables do not add much value since the slope of the linear relationship with the outcome is nearly zero. On the other hand, variables X_3, X_6, X_8 look to have stong linear associations.


beta*X2+res
300
200
100
0
−100
220
230
240
250
X2
beta*X3+res
200
100
0
−100
−200
60
70
80
90
100
X3
beta*X4+res
0
−100
−200
−300
−400
210
230
250
X4
beta*X5+res
100
0
−100
−200
−300
60
70
80
90
X5



2. Multicollinearity is addressed with the scatterplot matrix and Pearson coefficients table shown below. It indicates high multicollinearity in the covariates, particularly those of the math course scores with a high of $r = 0.77$ between 9^{th} and 10^{th} grade.



3. Automatic variable selection is used to begin the elimination process of variables, especially those with high multicollinearity. No variables are forced to stay in or out of the model. The results are shown in the R output below.

```
## Subset selection object
## Call: regsubsets.formula(Y ~ factor(X1) + X2 + X3 + X4 + X5 + X6 +
##     X7 + X8, data = project)
## 8 Variables  (and intercept)
##               Forced in Forced out
## factor(X1)1     FALSE      FALSE
## X2              FALSE      FALSE
## X3              FALSE      FALSE
## X4              FALSE      FALSE
## X5              FALSE      FALSE
## X6              FALSE      FALSE
## X7              FALSE      FALSE
## X8              FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          factor(X1)1 X2  X3  X4  X5  X6  X7  X8
## 1  ( 1 ) " "         " " " " " " " " "*" " " " "
## 2  ( 1 ) " "         " " " " " " " " "*" " " "*"
## 3  ( 1 ) "*"          " " " " " " " " "*" " " "*"
## 4  ( 1 ) "*"          " " "*" " " " " "*" " " "*"
## 5  ( 1 ) "*"          " " "*" "*" " " "*" " " "*"
## 6  ( 1 ) "*"          "*" "*" "*" " " "*" " " "*"
## 7  ( 1 ) "*"          "*" "*" "*" " " "*" "*" "*"
## 8  ( 1 ) "*"          "*" "*" "*" "*" "*" "*" "*"
```

In a separate plots of all of the criteria, R^2_{adj}, Bayes Information Criterion (BIC), and Mallow's C_p statistic are prepared below.

```
##   (Intercept) factor(X1)1 X2 X3 X4 X5 X6 X7 X8 adjr2     cp     bic
## 1           1           0  0  0  0  0  1  0  0 0.345 10.729 -32.861
## 2           1           0  0  0  0  0  1  0  1 0.421 -0.345 -41.354
## 3           1           1  0  0  0  0  1  0  1 0.422  0.578 -37.938
## 4           1           1  0  1  0  0  1  0  1 0.421  1.747 -34.268
## 5           1           1  0  1  1  0  1  0  1 0.417  3.443 -30.025
## 6           1           1  1  1  1  0  1  0  1 0.411  5.265 -25.646
## 7           1           1  1  1  1  0  1  1  1 0.406  7.103 -21.250
## 8           1           1  1  1  1  1  1  1  1 0.400  9.000 -16.789
```

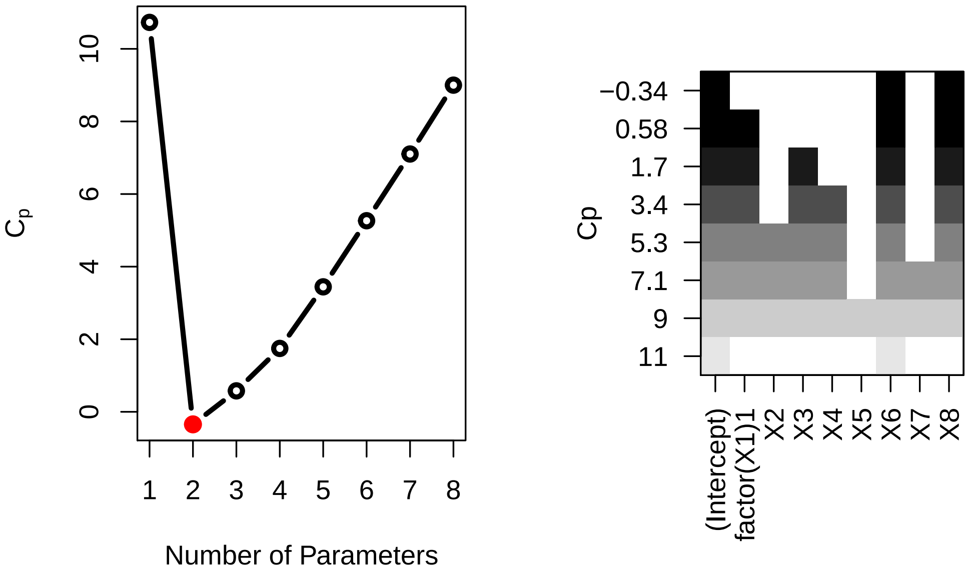


Figure 2: Cp variable Selection

With the best number of parameters highlighted in red in each plot, we determine that the models with two and three variables are the best overall. Variables X_6 and X_8 are in each of these models with X_1 being the third. These models will be compared to assess whether X_1 is a necessary covariate.

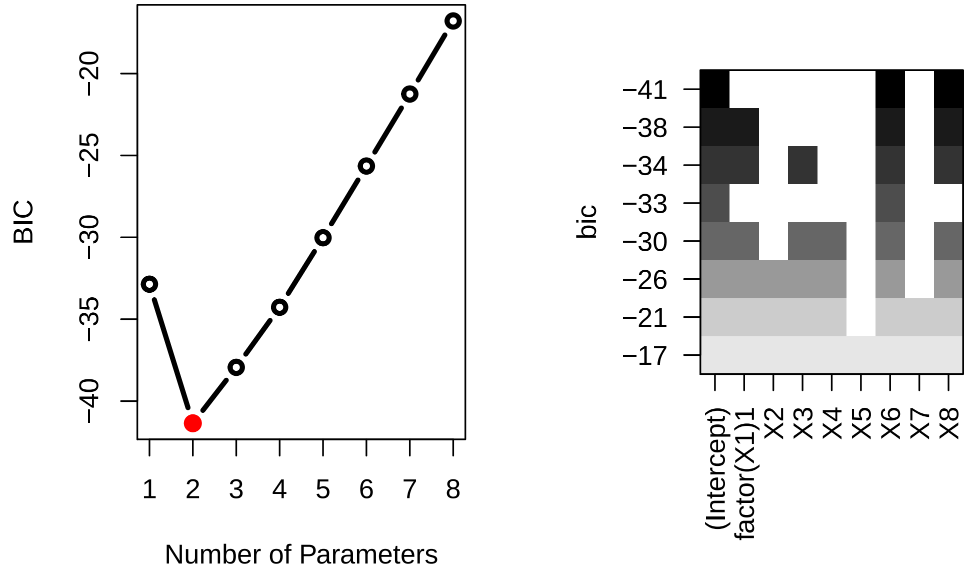


Figure 3: BIC variable Selection

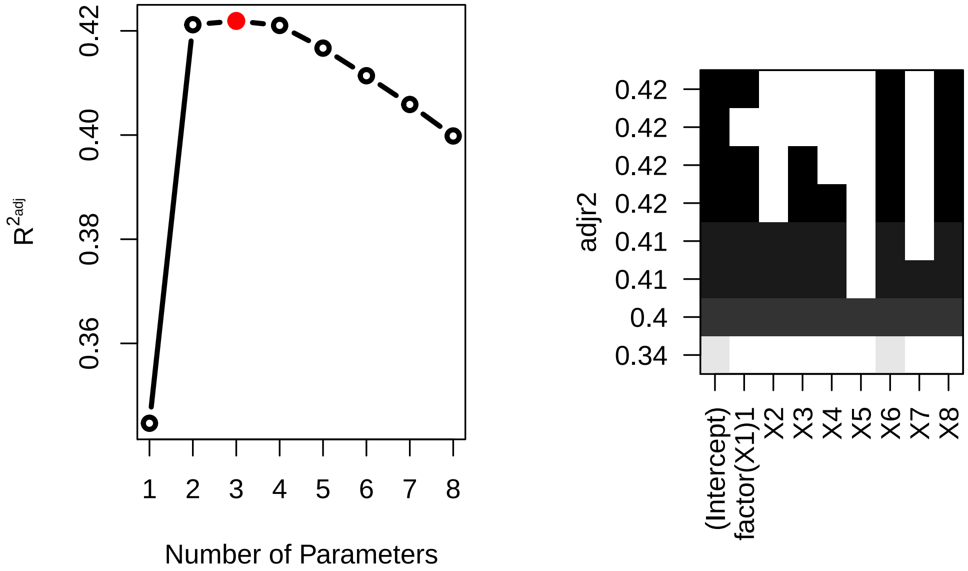


Figure 4: R squared adjusted variable Selection

4. Variable inflation factors are shown for the model $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_6 + \beta_3 X_8$ in the R output below. Since each inflation factor is less than the rule of thumb of 10, none

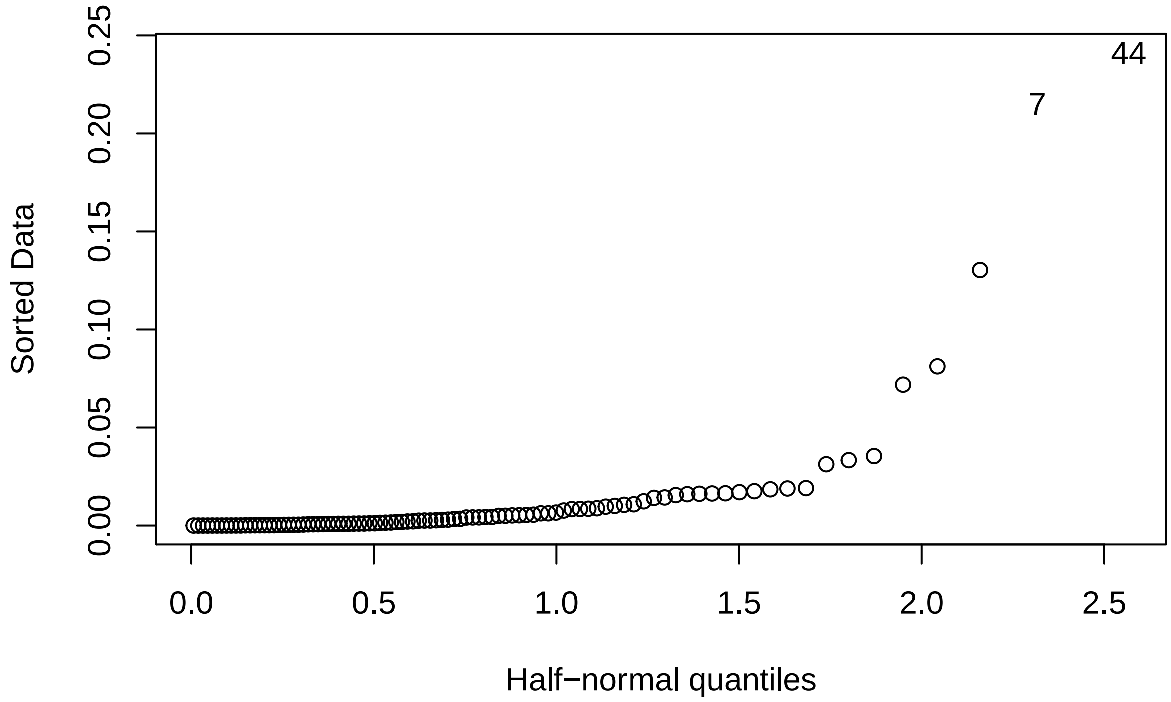


Figure 5: Cook's Distance in Half Normal Plot

present any immediate issues.

```
##       X1       X6       X8
## 1.887014 1.903293 1.681070
```

C. Residual Diagnostics

Potentially influential points are identified using Cook's distances and plotted with a half-normal plot. These unusually large or small values are shown in Figure 5 below and values of observations 7 and 44 are extracted, respectively.

```
##    X1  X6  X8   Y
## 18  1 278 100 580
## 73  0 229  79 260
```

A fit of the model without observations 7 and 44 shows that our model is changed drastically with the removal of these two points. For example, the adjusted R^2 increased from .422 to .490. Upon, investigating the source of these two observations (speaking to the administration of the school), the two students associated with the observations were considered abnormalities and could be removed from the investigation. Therefore, observations 7 and 44 are removed from the final model.

```
##
## Call:
## lm(formula = Y ~ X1 + X6 + X8, data = project_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.536  -35.770   -1.125   32.906  201.529
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -288.1986   162.8190  -1.770 0.080067 .
## X1             8.5368    15.0991   0.565 0.573202
## X6             2.5825     0.7172   3.601 0.000517 ***
## X8             2.9771     0.7090   4.199 6.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.43 on 91 degrees of freedom
## Multiple R-squared:  0.5058, Adjusted R-squared:  0.4895
## F-statistic: 31.04 on 3 and 91 DF,  p-value: 6.529e-14
```

Figure 6 shows the Studentized deleted residuals plotted against the expected values to validate the normality assumption. We can see that the residuals are nearly normal with one extreme value, identified to be the 8^{th} observation. A fit of the model without this observation displayed no change, therefore our model is considered robust to its inclusion.

A fit of the model without this observation displayed no change, therefore our model is considered robust to its inclusion.

```
##
## Call:
## lm(formula = Y ~ X1 + X6 + X8, data = project_test2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.423  -35.681   -1.735   33.207  201.999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -310.0552   164.1585  -1.889 0.062145 .
## X1             7.9180    15.1067   0.524 0.601470
## X6             2.6575     0.7208   3.687 0.000387 ***
## X8             3.0285     0.7106   4.262 4.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

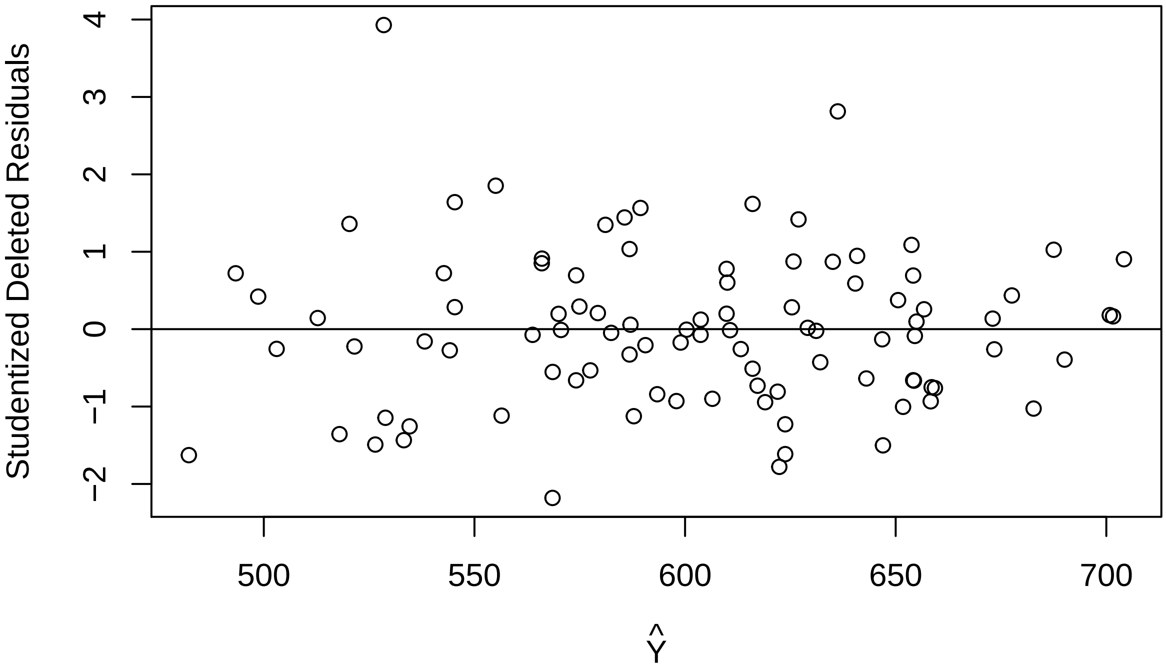


Figure 6: Studentized Residuals vs. Predicted SAT Values

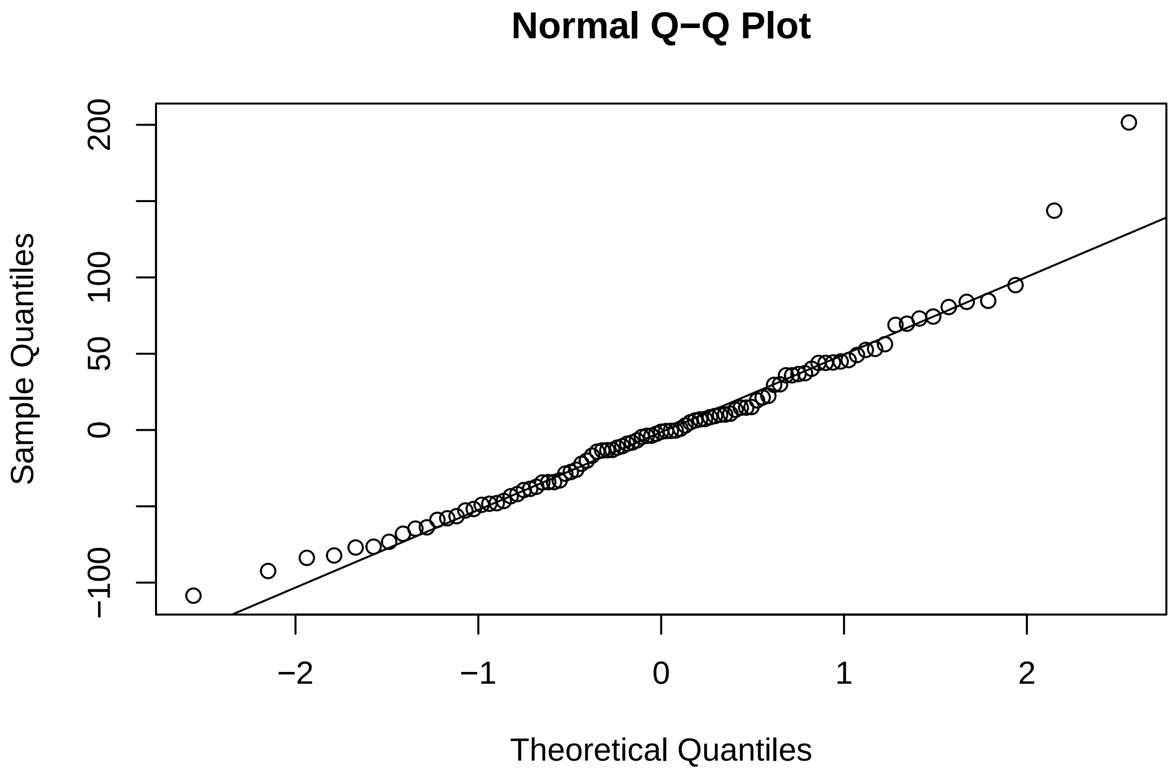


Figure 7: Q-Q Plot of Residuals

```
## Residual standard error: 52.41 on 90 degrees of freedom
## Multiple R-squared:  0.5107, Adjusted R-squared:  0.4944
## F-statistic: 31.32 on 3 and 90 DF,  p-value: 5.9e-14
```

To further support normality, a Q-Q plot is shown in Figure 7.